DOCUMENT RESUME

ED 447 207 TM 032 115

AUTHOR Witta, E. Lea
TITLE Comparison of Missing Data Treatments in Producing Factor Scores.
PUB DATE 2000-11-00
NOTE 17p.; Paper presented at the Annual Meeting of the American Evaluation Association (Honolulu, HI, November 1-5, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Methods; Factor Analysis; *Factor Structure; Likert Scales; Multivariate Analysis; Questionnaires; Regression (Statistics); *Responses; Surveys
IDENTIFIERS EM Algorithm; *Missing Data

ABSTRACT

Because ignoring the missing data in an evaluation may lead to results that are questionable, this study investigated the effects of use of four missing data handling techniques on a survey instrument. A questionnaire containing 35 5-point Likert-style questions was completed by 384 respondents. Of these, 166 (43%) questionnaires contained 1 or more missing responses. The missing data pattern was non-ignorable. Listwise deletion, pairwise deletion, regression, and the expectation maximization (EM) algorithm were used to treat the missing data. Resulting data were then submitted to factor analysis and factor scores obtained. Factor scores for each group defined by missing data method were then contrasted by multivariate analysis of variance. Less than 1% of the variance in scores could be explained by group (F=0.218, df=30,3496, p=1.0). Based on these results, the choice of missing value treatment can be based on the consequences of loss of power by loss of cases or other data handling considerations. (Contains 1 table, 1 figure, and 17 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 447 207

Running Head: Factor Score by Missing Data Treatment

Comparison of Missing Data Treatments in Producing Factor Scores

E. Lea Witta

The University of Central Florida

Department of Educational Foundations

TM032115

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

E. Witta

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper presented at the annual conference of the American Evaluation Association, Honolulu, Hawaii, November 1-5, 2000. For further information contact Lea Witta lwitta@mail.ucf.edu

BEST COPY AVAILABLE

Abstract

Because ignoring the missing data in an evaluation may lead to results that are questionable, this study investigated the effects of use of four missing data handling techniques on a survey instrument. A questionnaire containing 35 five-point Likert style questions was completed by 384 respondents. Of these, 166 (43%) questionnaires contained one or more missing responses. The missing data pattern was non-ignorable. Listwise deletion, pairwise deletion, regression, and the expectation maximization algorithm were used to treat the missing data. Resulting data was then submitted to factor analysis and factor scores obtained. Factor scores for each group defined by missing data method were then contrasted by multivariate analysis of variance. Less than 1% of the variance in scores could be explained by group($F=.218$, $df=30, 3496$, $p=1.0$). What if analyses were conducted by reducing the number of complete responses.

When conducting survey research, often the data analyzed contains missing values. If the cause of the missing values is known, the solution to this problem may be incorporated in the study. Seemingly unavoidably, however, survey subjects fail to answer some questions for unknown reasons. Ignoring this problem may threaten the quality of the study and data analysis.

In addition, several researchers have shown that different methods of handling missing values may produce different results. Witta and Kaiser (1991) reported that the regression coefficients and total variance accounted for by the variables changed depending on the method used to handle missing values. After re-analyzing three studies of private/public school achievement, Ward and Clark III (1991) concluded that the method used to handle missing data influenced the outcome of these studies. Mundfrom and Whitcomb (1998) found differences in correct classification of subjects based upon the missing data handling procedure used.

Statement of the Problem

The purpose of the current study was to examine the influence of missing data methods on factors created from a survey of high school student participants in an educational interactive video system. The missing data methods studied were listwise deletion, pairwise deletion, regression and expectation maximization. The sample consisted of 384 respondents. Of these, 166 (43%) of the surveys contained one or more missing values.

Until recently, the only methods available with popular statistical computer software focused on handling the missing data problem by deleting subjects with incomplete information, deleting the variables with missing values, or replacing the missing value with some reasonable estimate. Now, however, new subroutines are available to provide more assistance in handling missing data and providing analysis choices using iterative regression or expectation maximization

(EM) procedures. These relative new methods (in current software) also provide the possibility of specifying the model to be used (i.e., multivariate normality, adding a randomly selected error).

Review of Literature

Listwise Deletion

Listwise deletion is probably the most frequently used method of handling missing data. This is usually the default option in statistical software programs. This method discards cases with a missing value on any variable and thus is very wasteful of data. Listwise deletion, however, has been shown to be effective with low average intercorrelation, less than four variables and a small proportion of missing values (Chan, et.al., 1976; Haitovsky, 1968; Timm, 1970). The assumption of missing completely at random is crucial to the use of this method. It is more likely, however, to find the complete sample different in important ways from the incomplete sample (Little & Rubin, 1987). Problems for a researcher using this method include a reduction in power, an increase in standard error, and the elimination of sub-populations.

Pairwise Deletion

When using pairwise deletion, covariances are computed between all pairs of variables having both observations, eliminating those that have a missing value for one of the two variables (Glasser, 1964). Means and variances are computed on all available observations. It is assumed that the use of the maximum number of pairs and all the individual observations yield more valid estimates of the relationship between the variables. This method is sometimes referred to as a "complete data" method. It is also assumed that when two variables are correlated, information on one improves the estimates of the other variable. An additional assumption is that the pairs are a random subset of the sample pairs. If these assumptions are true, pairwise deletion produces

unbiased estimates of the variable means and variances (Hertel, 1976). When missing data are not missing completely at random, however, the correlation matrix produced by pairwise deletion may not be Gramian (Norusis, 1988). When Marsh (1998) investigated the estimates produced by pairwise deletion for randomly missing data, heconcluded parameter variability was explained, parameter estimates were unbiased, and only one covariance matrix was nonpositive definite.

Regression

The regression methods of handling missing data rely on information contained in non-missing values of other variables to provide estimates of missing values. As the correlation between variables and the number of variables increases, the regression methods, theoretically, perform better. Too many variables, however, can cause problems with over prediction (Kaiser & Tracy, 1988) and too high an average intercorrelation can result in a singular matrix. In these cases, regression does not perform well.

Variations in the regression methods include differences in methods of developing the initial correlation matrix (listwise deletion, pairwise deletion, and mean substitution) and the presence or absence of iteration procedures. Differences in regression methods also include the use of randomly selected residuals for iterations and assumptions of a normal distribution. Mundfrom and Whitcomb (1998) investigated the effects of using mean substitution, hot-deck imputation, and regression imputation on classification of cardiac patients. Mean substitution and hot-deck imputation correctly classified patients more frequently than regression imputation.

Expectation Maximization

Dempster, Laird, and Rubin (1977) recommended the use of the EM algorithm which imputes estimates simultaneously in an iterative procedure. The E step of this algorithm finds the

conditional expectation of the missing values. The M step performs maximum likelihood estimation as if there were no missing data. The primary difference between this procedure and the regression procedure is that the values for the missing data are not imputed and then iterated. The missing values are functions based on the conditional expectation (Little & Rubin, 1987). This method of handling missing data represents a fundamental shift in the way of thinking about missing data (Schafer & Olsen, 1998).

Pattern of Missing Values

All of the missing data handling procedures discussed require data missing at random (MAR) or missing completely at random (MCAR). Missing completely at random as described by Little and Rubin (1987) means that the probability of missingness is independent of that variable's value and the value of any other variable in the data set. Missing at random as described by Little and Rubin (1987) means that the probability of missingness may depend on another variable, but not on the value of the variable itself. Yet Cohen and Cohen (1983) suggested that in survey research the absence of data on one variable may be related to another variable and may be due to the value of the variable itself. When investigating simultaneously missing values, Witta (1996/97) found concurrently missing values ($p<.001$) in three of four samples using data from a national database and Witta (2000a) found concurrently missing values ($p<.001$) in all four samples using data from a national database. Little's test of missing completely at random is now included with the missing data subroutine of SPSS. Unfortunately, because missing at random requires knowledge of the true value of the missing data, "there is no magic test for MAR" (Hill, 1997, p. 43). Schafer and Olsen (1998), however, argue convincingly that "every missing-data method must make some largely untestable statistical assumptions about the manner in which the missing

values were lost" (p551). Consequently, when analyzing real data, researchers typically assume missing at random.

Method

Subjects

All high school students enrolled in an interactive video class at this facility during the Spring semester, 1998, were surveyed. Questionnaires containing 35 five-point Likert style questions, some demographic questions, and some open-ened questions were administered during the regularly scheduled class time by the class instructor or remote facilitator. This analysis used only the 35 Likert type questions.

Of the 384 returned surveys, respondents represented 19 classes and 24 high schools. Fifty-four percent of the sample were female and 59% were from the home site. One hundred sixty-six (43%) questionnaires contained one or more missing responses. Regression and the expectation maximization algorithm were used to treat the missing data using the missing data subroutine 7.5 in SPSS 10.0. The resultant data for each procedure was saved in a new file. In addition, the pattern of missing data, and Little's test of missing completely at random was recorded. The original data was replicated to create a data file for pairwise deletion analysis.

Each of the four data sets was then submitted to principal components analysis in SPSS 10.0. The factor scores produced by varimax rotation were saved with the data set. Finally, the four data sets were merged and factor scores for each group defined by missing data method were contrasted by multivariate analysis of variance.

Results

When data were examined in the missing data procedure in SPSS, the test of missing

completely at random was statistically significant ($\chi^2 = 3292.352$, df = 2910, $p < .001$). To examine the pattern of missing values, individual categories were created for any variable having 4 or cases with missing values. All other variables containing a missing value, but having less than 4 (<1%) cases with a missing value were categorized as "Other". Over half of the incomplete cases were classified as "Other" or produced by missing values on an assortment of variables as is depicted in Figure 1. Thus, visual examination of the pattern of missing values did not reveal any obvious patterns of missing values.

Insert Figure 1 About Here

After treatment by a missing data method, the resultant data sets from the EM algorithm and and regression were analyzed using principal components analysis (PCA) with varimax rotation. The pairwise and listwise deletion data sets were analyzed in the same procedure using the pairwise and listwise missing setting in PCA. Using eigenvalues of 1 or larger, all data sets produced 10 factors. Although there were differences in the order in which factors were listed and, consequently, the variance explained by each, the factors produced were essentially the same. In addition, the factor loadings were very similar as is depicted in Table 1.

Insert Table 1 About Here

The mean vector of factor scores produced by each missing data handling treatment were then contrasted by multivariate analysis in SPSS 10.0. Less than 1% of the variance in the mean

vector of factor scores could be explained by group (F=.218, df=30, 3496, p=1.0).

To provide an estimate of what if 50% of the cases contained missing values, 166 complete cases were randomly sampled from the 218 complete cases and merged with the 166 incomplete cases providing a data set of 332 cases. The previous procedure (treatment by missing data method, principal components analysis, varimax rotation, and creation of factor scores) was repeated. When tested for missing completely at random Little's MCAR test (χ^2 = 3303.005, df = 2994, $p < .001$) again yielded statistically significant results. As anticipated the cases were not missing completely at random. When the factor mean vectors were contrasted, less than 1% of the variance in scores was explained by missing data treatment group (F=.236, df 30, 2945, p=1.0).

For the final what-if analysis, 100 complete cases were randomly sampled from the 218 complete cases and merged with the 166 incomplete cases providing a data set of 266 cases. Missing data treatments were again applied and factor scores saved. Little's MCAR test(χ^2 = 3176.278, df = 2994, p = .010) was once more statistically significant. In this instance, however, only 9 factors emerged from the data. Again less than 1% of the variance in the mean vector of factor scores could be explained by missing data treatment group (F=.26, df=27, 2103, p=1.0).

Discussion and Conclusion

In all cases used in this study there were no statistically significant differences in factor score mean vectors produced by the missing value treatments even though the pattern of missing values was not missing completely at random. In addition, the differences by missing value treatment group explained less than 1% of the variance in mean vectors. Based on these results, the choice of missing value treatment can be based upon consequences of loss of power by loss of cases or other data handling considerations.

There was, however, a difference in the pattern of missing values in this study and those observed in other studies. In a study of missing values using a sample from the NELS 88 (National Educational Longitudinal Study of 1988), Witta (2000b) detected differences in missing data handling treatments. The pattern of missing values in that study was also not missing completely at random. The observed pattern, however, differed from the current study. In Witta's (2000b) study, the category "Other" was less than 35% of the incomplete cases under all conditions. A few variables, specifically standardized test score, accounted for the remaining 65% of the cases. In the current study, the category "Other" represented more than 50% of the incomplete cases. This category consisted of variables with less than 1% of the values missing. It appears, therefore, that the effectiveness of the missing value methods are seriously affected by the pattern of missing values. Because there is no test for 'Missing at Random', an assumption is made that this condition is satisfied when testing effectiveness of missing data method. In the current study, this assumption appears to be satisfied. In Witta's (2000b) study, that assumption appeared to be violated. As a result, it is strongly recommenced that all studies first investigate the pattern of missing values. It is further suggested that future research investigate the patterns of missing values in an attempt to determine when it would be acceptable to assume 'Missing at Random'.

References

Chan, L.S., Gilman, J.A., & Dunn, O.J. (1976). Alternative approaches to missing values in discriminant analysis. *Journal of the American Statistical Association, 71*, 842-844.

Dempster, A.P., Laird, N.W., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B, 39*, 1-38.

Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association, 59*, 834-844.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journalof the Royal Statistical Society, B, 30*, 67-82.

Hill, M.A. (1997). *SPSS Missing Value Analysis 7.5* [Computer program manual]. Chicago: SPSS Inc.

Kaiser, J. & Tracy, D.B. (1988). Estimation of missing values by predicted score. *Proceedings of the Section on Survey Research, American Statistical Association 1988*. 631-635.

Little, R.J.A., & Rubin, D.R. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Marsh, H.W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling, 5* (1), p 22-36.

Mundfrom, D.J. & Whitcomb, A. (1998). *Imputing missing values: The effect on the accuracy of classification*. Paper presented at the annual meeting of the American Educational Research Association, San Diego. ED419817.

Norusis, M.J. (1988). *SPSS-X Introductory Statistics Guide: Release 3* [Computer program manual]. (pp 107-108). Chicago: SPSS Inc.

Schafer, J.L. & Olsen, M.K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavior Research, 33* (4), p 545-571.

Timm, N.H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika, 35*, 417-437.

Ward, Jr., T.J. & Clark III, H.T. (1991). A reexamination of public-versus private-school achievement: the case for missing data. *Journal of Educational Research, 84*, 153-163.

Witta, E.L. (2000a). Are values missing randomly in survey research? *Educational Research Quarterly, 23* (3), pp 44-56.

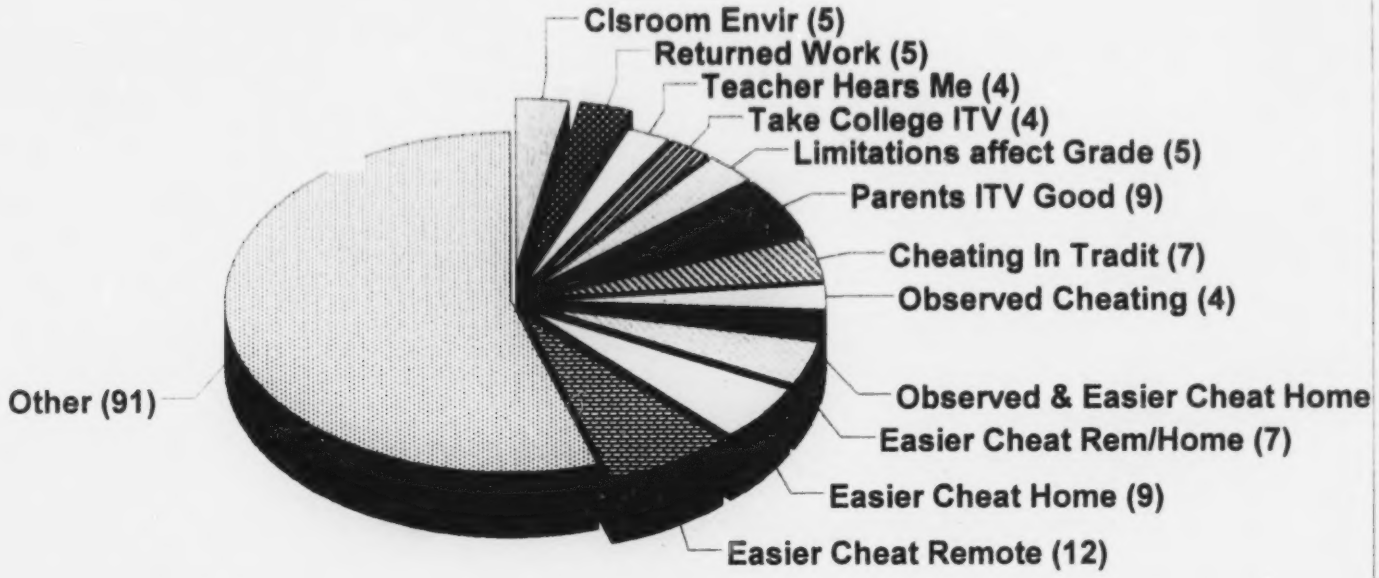
Witta, E. L. (2000b). *Four methods of handling missing data in predicting educational achievement* (ERIC Document Reproduction Service, ED to be assigned - Tracking number **TM031627**).

Witta, E.L. (1996/97). Randomness of missing values in survey data. *Louisiana Education Research Journal, XXII* (2), p 73-86.

Witta L. & Kaiser, J. (1991, November). *Four methods of handling missing data with GSS-84*. Paper presented at the meeting of the Mid-South Educational Research Association, Lexington, KY

Pattern of Missing Values

High School Students



Little's MCAR test: $\chi^2 = 3292.352$, df = 2910, Prob = .000

Table 1

Factor Loadings By Missing Data Treatment

Question	Pairwise Deletion	EM Algorithym	Regression	Listwise Deletion
Factor	1	2	2	1
Q26	0.80	0.80	0.80	0.81
Q25	0.77	0.76	0.76	0.83
Q28	0.76	0.75	0.75	0.78
Q24	0.75	0.75	0.75	0.76
Q23	0.68	0.70	0.68	0.66
Q29		0.48	0.49	
Q5		0.45	0.45	
Q6				
Factor	3	3	3	3
Q32	0.82	0.83	0.83	0.81
Q21	0.79	0.80	0.78	0.74
Q13	0.73	0.72	0.72	0.78
Q30		0.50	0.46	
Factor	5	5	5	4
Q10	0.99	0.99	0.97	0.98
Q11	0.99	0.99	0.98	0.98
Factor	7	7	7	9
Q27	0.78	0.78	0.78	0.72
Q31	0.71	0.71	0.70	0.66
Factor	9	9	9	7
Q9	0.77	0.77	0.76	0.70
Q8	0.56	0.57	0.56	0.50
Q4				0.48
Q5				0.43

Question	Pairwise Deletion	EM Algorithym	Regression	Listwise Deletion
Factor	2	1	1	2
Q17	0.82	0.82	0.82	0.82
Q18	0.77	0.79	0.78	0.78
Q16	0.75	0.75	0.76	0.77
Q7	0.75	0.76	0.75	0.78
Q14	0.73	0.73	0.72	0.73
Q15	0.63	0.63	0.62	0.52
Q33	0.56	0.56	0.58	0.60
Factor	4	4	4	5
Q2	0.69	0.69	0.70	0.69
Q3	0.64	0.65	0.65	0.62
Q4	0.61	0.60	0.61	
Q1	0.54	0.56	0.55	0.57
Factor	6	6	6	6
Q35	0.86	0.85	0.85	0.81
Q34	0.75	0.75	0.76	0.71
Factor	8	8	8	8
Q12	0.75	0.75	0.73	0.63
Q22	0.64	0.61	0.64	0.72
Q30				0.43
Factor	10	10	10	10
Q20	0.83	0.82	0.81	0.84
Q19	0.65	0.65	0.65	0.59
Q6		0.42	0.46	

Note. Loadings less than 0.40 were blanked.